A new method for deriving waiting-time approximations in polling systems with renewal arrivals

J.L. Dorsman[1,2], R.D. van der Mei[1,3] and E.M.M. Winands[3]

[1]CWI, Probability and Stochastic Networks, Amsterdam, The Netherlands
[2]Eindhoven University of Technology, EURANDOM, Eindhoven, The Netherlands
[3]VU University Amsterdam, Department of Mathematics, Amsterdam, The Netherlands
j.l.dorsman@tue.nl, R.D.van.der.Mei@cwi.nl, emm.winands@few.vu.nl

Abstract

We study the waiting-time distributions in cyclic polling models with renewal arrivals, general service and switch-over times, and exhaustive service at each of the queues. The assumption of renewal arrivals prohibits an exact analysis and reduces the available analytic results to heavy-traffic asymptotics, limiting results for large switch-over times and large numbers of queues, and some numerical algorithms. Motivated by this, the goal of this paper is to propose a new method for deriving simple closed-form approximations for the complete waiting-time distributions that work well for arbitrary load values. Extensive simulation results show that the approximations are highly accurate over a wide range of parameter settings.

Key words: Polling systems, renewal arrivals, waiting-time distribution, approximation

1. Introduction

Polling systems are queueing systems consisting of multiple queues, attended by a single server that visits the queues to serve waiting customers. Whenever the server proceeds from one queue to another, a switch-over time is incurred. Typically, the queues are visited in a cyclic order. Polling models occur naturally in the modelling of applications where different types of customers compete for access to a common resource. Typical application areas of polling models are computer-communication systems, manufacturing

systems, maintenance systems and traffic systems. We refer to [9] for an overview of the applicability of polling models, and to [15, 17] for overviews of the available results. In the present paper, we study cyclic polling with exhaustive service at all queues and with renewal arrival processes.

In the literature, the vast majority of papers on polling models rely on the assumption of Poisson arrivals, whereas for models with non-Poisson arrivals hardly any exact results on the waiting-time distributions are known. The most general results known for renewal-driven polling models are asymptotic results, such as heavy-traffic asymptotics [13] or asymptotics for large switch-over times [19]. Closed-form approximations are available for the mean waiting-time [2], but today there is no simple approximation available for the tail probabilities of the waiting times at each of the queues. Even in case of Poisson arrivals, generally little attention is paid to the tail probabilities of the waiting-time distributions. As an exception, assuming Poisson arrivals, Choudhury and Whitt [3] propose an efficient numerical algorithm to calculate the moments and the tail probabilities of the waiting times based on numerical transform inversion, for branching-type polling models [14]. For models that violate the branching structure, more computationally intensive algorithms exist [1, 8]. A common drawback of these numerical algorithms is that they only give limited insight into how the waiting-time distribution reacts to changes in the system parameters. These observations raise the need for the development of simple yet accurate approximations for the tail probabilities of the waiting times for polling models with renewal arrivals.

Motivated by this, in this paper we propose a new method for deriving closed-form approximations of the waiting-time distribution for arbitrary values of the load. The approach taken is that of combining known heavy-traffic (HT) asymptotics for the waiting-time distributions [11, 13], which work well when the system is heavily loaded, with a recently developed approximation for the mean waiting times in [2], which works well for the whole range of load values. This paper presents the first *closed-form* approximation of the waiting-time *distribution* in polling systems with renewal arrivals. In this regard, the present research is an extension of [2], which derives approximate expressions for the *mean* waiting times. This approximation is shown to be exact in the known limiting cases, and extensive experimentation with simulations shows that the approximation is highly accurate for a wide range of parameter settings. We emphasize that the strength of this combined

approach lies in its striking simplicity and the fact that it leads to approximations in closed form, which opens up many possibilities for generalization of the approach to other polling models (e.g., with more general branching-type service policies, and with non-cyclic periodic server routing) and for optimization of the system performance with respect to the system parameters.

The remainder of this paper is organized as follows. In Section 2 we introduce the model and notation. In Section 3 we present the distributional approximation, which is the main result of this paper. In Section 4 we discuss properties of the approximation, and in Section 5, the accuracy of the approximation is evaluated by an extensive simulation study. Finally, in Section 6 we discuss suggestions for further research.

2. Model description and notation

Consider a polling system consisting of $N \geq 1$ queues, $Q_1, \ldots, Q_N$, with an infinite-sized buffer. Customers arrive at Q_i according to a renewal process, with rate $\lambda_i = \frac{1}{\mathbb{E}[A_i]}$, where A_i is the random variable describing the interarrival times of customers at Q_i. The total arrival rate to the system is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. Within a queue, customers are served on a first-in-first-out (FIFO) basis. The service time of a type-i customer at Q_i is denoted by the random variable B_i with k^{th} moment $\mathbb{E}[B_i^k]$, and its waiting time in Q_i by the random variable W_i with k^{th} moment $\mathbb{E}[W_i^k]$, $k > 0$. The random variable B denotes the service time of an arbitrary customer entering the system, with $\mathbb{E}[B^k] = \sum_{i=1}^{N} \frac{\lambda_i}{\Lambda} \mathbb{E}[B_i^k]$. Queues are served according to an *exhaustive* service discipline, and as soon as Q_i becomes empty, the server proceeds to Q_{i+1}. We define a cycle at Q_i as the time between two successive departures of the server from Q_i. The time needed by the server to switch from Q_i to Q_{i+1} is denoted by the random variable S_i with k^{th} moment $\mathbb{E}[S_i^k]$ $(k > 0)$. Let the random variable $S = \sum_{i=1}^{N} S_i$ denote the total switch-over time in a cycle. Throughout, it is assumed that $\mathbb{E}[S] > 0$ and that all inter-arrival times, service times and switch-over times are mutually independent and independent of the state of the system. The load offered to Q_i is denoted by $\rho_i = \lambda_i \mathbb{E}[B_i]$, $1 \leq i \leq N$. The total load in the system is denoted by $\rho = \sum_{i=1}^{N} \rho_i > 0$. A necessary and sufficient condition for the stability of the described system is $\rho < 1$ [5]. The waiting time at Q_i is defined as the time between the arrival of an arbitrary customer in the

system and the moment when he is taken into service.

Throughout, it may be convenient to scale a system such that a certain load is achieved. This scaling is done by keeping the service time distributions fixed and varying the rates of the renewal processes. In particular, it proves convenient to denote with $\hat{x}$ the value of each variable x that is a function of ρ evaluated at $\rho = 1$. $\hat{A}_i$ then denotes the inter-arrival time of Q_i customers evaluated at $\rho = 1$. Then, scaling to a load $\rho < 1$ is done by taking the random variable $A_i := \hat{A}_i/\rho$.

Finally, we introduce some notation. The residual length of a random variable X is denoted by X^{res}, with $\mathbb{E}[X^{res}] = \frac{\mathbb{E}[X^2]}{2\mathbb{E}[X]}$. The squared coefficient of variation (SCV) of a random variable X, $\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$ is denoted by c_X^2. We define $\sigma^2 = \sum_{i=1}^N \hat{\lambda}_i(\mathrm{Var}[B_i] + c_{A_i}^2\mathbb{E}[B_i]^2)$ and $\delta = \sum_{j=1}^N \sum_{k=j+1}^N \hat{\rho}_j\hat{\rho}_k$; note that in the case of Poisson arrivals, the former can be simplified to $\sigma^2 = \mathbb{E}[B^2]/\mathbb{E}[B]$. The notation $\xrightarrow{d}$ means convergence in distribution, and $\mathbf{1}_{\{A\}}$ denotes the indicator function on the event A. Finally, when a random variable X is said to have a gamma distribution with shape parameter α and inverse scale parameter μ, its density function is given by $f_X(x) = e^{-\mu x}\mu^\alpha x^{\alpha-1}\mathbf{1}_{\{x\geq 0\}}/\Gamma(\alpha)$, where $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$.

3. Derivation of the approximation

The two key ingredients of the distributional approximation will be the HT diffusion approximation for the waiting-time by Olsen and Van der Mei [13], and the mean waiting-time approximation by Boon et al. [2] for a general values of the load $\rho < 1$. The HT diffusion approximation will be refined such that its mean concides with the mean waiting-time approximation, while the diffusion approximation remains unchanged in the case of HT after refinement. The two ingredients are given first, after which the main result is derived and presented. Although not stated explicitly, it follows naturally from [13] that

$$(1-\rho)W_i \xrightarrow{d} UI_i, \quad \rho \uparrow 1, \tag{1}$$

where U is a uniformly distributed random variable on [0,1], and I_i a gamma distributed random variable with parameters

$$\alpha = \frac{2\mathbb{E}[S]\delta}{\sigma^2} + 1 \text{ and } \mu_i = \frac{2\delta}{(1-\hat{\rho}_i)\sigma^2}. \tag{2}$$

Let $\mathbb{E}[W_i]$ denote the mean waiting-time of a type-i customer at Q_i, $1 \leq i \leq N$. Then, Boon et al. [2] propose the following approximation $\mathbb{E}[W_{i,Boon}]$ for the mean waiting times:

$$\mathbb{E}[W_{i,Boon}] = \frac{K_0 + K_{1,i}\rho + K_{2,i}\rho^2}{1-\rho}, \tag{3}$$

where the constants $K_0, K_{1,i}$ and $K_{2,i}$ depend on several parameters of the polling system at hand:

$$\begin{aligned}
K_0 &= \mathbb{E}[S^{res}], \\
K_{1,i} &= \hat{\rho}_i((c_{A_i}^2)^4 \mathbf{1}_{\{c_{A_i}^2 \leq 1\}} + 2\frac{c_{A_i}^2}{c_{A_i}^2+1}\mathbf{1}_{\{c_{A_i}^2 > 1\}} - 1)\mathbb{E}[B_i^{res}] + \mathbb{E}[B^{res}] \\
&\quad + \hat{\rho}_i(\mathbb{E}[S^{res}] - \mathbb{E}[S]) - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathrm{Var}[S_{i+j}], \\
K_{2,i} &= \frac{1-\hat{\rho}_i}{2}\Big(\frac{\sigma^2}{2\delta} + \mathbb{E}[S]\Big) - K_0 - K_{1,i}.
\end{aligned}$$

In fact, $\mathbb{E}[W_{i,Boon}]$ is an interpolation between a light-traffic (LT) limit and a HT limit of the mean waiting-time; K_0 and $K_{1,i}$ stem from the LT limit, while $K_{2,i}$ is based directly on the HT limit.

Our distributional approximation will be a refinement of the HT diffusion approximation given in (1). To this end, in the spirit of (1) we assume that the waiting-time distribution of Q_i can be well approximated by a product of a uniform random variable on [0,1] and a gamma random variable with α_{ia} and μ_{ia}, divided by $(1-\rho)$. To parameterize α_{ia} and μ_{ia}, we formulate the following three requirements for the refinement of the approximation in (3):

1. *In HT the refined approximation must coincide with the diffusion approximation of* [13]*, i.e.,*

$$\frac{\alpha}{\alpha_{ia}} \to 1 \;\; \textit{and} \;\; \frac{\mu_i}{\mu_{ia}} \to 1, \;\; \textit{when} \;\; \rho \uparrow 1.$$

2. *The expectation of the refined approximation coincides with the mean waiting-time approximation of* [2].

3. *The SCV of the refined approximating distribution matches the SCV of the HT diffusion approximation by* [13]*, so that the shape of the refined diffusion approximation matches the shape of the HT diffusion approximation.*

These three requirements uniquely determine the parameters α_{ia} and μ_{ia}, leading to the following approximation for the waiting-time distribution in polling systems with renewal arrivals and $\rho < 1$:

$$\mathbb{P}[W_i < x] \approx \mathbb{P}[UI_{i,app} < (1-\rho)x], \tag{4}$$

where U is a uniformly distributed random variable on [0,1], and $I_{i,app}$ a gamma distribution with parameters

$$\alpha_{ia} = \alpha_a = \frac{2\mathbb{E}[S]\delta}{\sigma^2} + 1 \text{ and } \mu_{ia} = \frac{2\mathbb{E}[S]\delta + \sigma^2}{2\sigma^2(1-\rho)\mathbb{E}[W_{i,Boon}]}. \tag{5}$$

It can be verified that the k^{th} moment of the obtained distributional approximation can be expressed as follows, for $k \geq 1$,

$$\mathbb{E}[W_{i,app}^k] = \frac{1}{(1-\rho)^k}\frac{1}{k+1}\prod_{i=0}^{k-1}\frac{\alpha_a + i}{\mu_{ia}} = \frac{2^k\mathbb{E}[W_{i,Boon}]^k}{k+1}\prod_{i=1}^{k}\frac{2\mathbb{E}[S]\delta + i\sigma^2}{2\mathbb{E}[S]\delta + \sigma^2}, \tag{6}$$

with α_a and μ_{ia} as defined above.

Note that we have tried a number of adjustments, but the one presented in the third requirement turned out to be the most robust and intuitively appealing (see also Remark 3). Other natural possibilities for adjustment are, for example, matching the variance of the HT diffusion approximation

or an interpolation between the SCV in LT and HT.

We end this section with a number of remarks.

Remark 1 (Olsen's approximation). A refined diffusion approximation for the distribution of the waiting time in polling systems with Poisson arrivals was presented by Olsen [12]. The HT diffusion approximation used in [12] consists of a uniformly distributed random variable on [0,1] "times" a gamma distribution with shape parameter $\frac{\mathbb{E}[S]\sum_{i=1}^{N}\rho_i(\rho-\rho_i)}{\sum_{i=1}^{N}\lambda_i(\mathrm{Var}[B_i]+\mathbb{E}[B_i]^2)}+1$ and inverse scale parameter $\frac{(1-\rho)\sum_{i=1}^{N}\rho_i(\rho-\rho_i)}{(1-\rho_i)\sum_{i=1}^{N}\lambda_i(\mathrm{Var}[B_i]+\mathbb{E}[B_i]^2)}$. Note that in HT these parameters coincide with the ones used in (1). Refinement is done using an approximation of the mean delay obtained by [6] for Poisson arrivals. Suggested by this mean delay approximation, Olsen adds an extra factor of ρ in the shape parameter, such that it becomes $\frac{\mathbb{E}[S]\sum_{i=1}^{N}\rho_i(\rho-\rho_i)}{\rho\sum_{i=1}^{N}\lambda_i(\mathrm{Var}[B_i]+\mathbb{E}[B_i]^2)}+1=\frac{2\mathbb{E}[S]\delta}{\sigma^2}+1$. The inverse scale parameter is changed accordingly such that the approximation satisfies the mean delay approximation in [6]. One can verify that in case of Poisson arrivals, the shape parameters of Olsen's approximation and our approximation coincide. Hence, the distributional approximation as given in this paper generalises Olsen's approximation to systems with renewal arrivals, and the presented derivation of the main result of this paper creates intuition and justification behind the distributional approximation of [12].

Remark 2 (Information availability). The derived waiting-time distribution approximation (4) only requires the first two moments of the interarrival, service and switch-over time distributions as an input, whereas the complete waiting-time distribution generally depends on their complete distributions, even for Poisson arrivals. This makes the approximations useful for practical purposes, because in reality information about more than the first two moments is often hard to get.

Remark 3 (Applicability). Yet another view is provided by the notion that the derived approximation gives a procedure to estimate the complete waiting-time distribution based on the mean waiting time and aggregate measures for imbalance δ and variability σ^2. In this regard, it is important to note that the mean waiting-time can easily be measured in real-life applications, in contrast to higher moments or tail probabilities.

Remark 4 (Other service disciplines). Given that the two key ingredients are available for polling systems with other service disciplines, distributational waiting-time approximations can be derived for these classes of polling systems as well using the method as described above. To illustrate this, consider a polling model where all queues are served according to a *gated* service discipline, i.e., where during a visiting period at a queue the server will only serve the customers which arrived before the start of this period. For this service discipline, it follows from [13] that

$$(1-\rho)W_i \xrightarrow{d} UI_i, \quad \rho \uparrow 1, \tag{7}$$

where U is uniform on $[\hat{\rho}_i, 1]$, and where I_i is a gamma distribution with parameters α and $(1-\hat{\rho}_i)\mu_i$, with α and μ_i as given in (2). As for the second key ingredient, [2] also contains an approximation of the delay's first moment for the gated service discipline. This approximation still has the form of (3), however with different values for $K_0, K_{1,i}$ and $K_{2,i}$, derived in [2]. Using these two key ingredients, one obtains for the distributional approximation

$$\mathbb{P}[W_i < x] \approx \mathbb{P}[UI_{i,app} < (1-\rho)x], \tag{8}$$

where U is uniform on $[\hat{\rho}_i, 1]$ and where $I_{i,app}$ has a gamma distribution with parameters α_a and $(1+\hat{\rho}_i)\mu_{ia}$, with α_a and μ_{ia} as given in (5). Note that $\mathbb{E}[W_{i,Boon}]$ in the expression of μ_{ia} in this case refers to the gated version of Boon's approximation, derived in [2].

4. Alignment with asymptotic regimes

In [2] it is shown that the first moment of the distributional approximation is in line with several known exact results, which gives support for the quality of the approximation. For Poisson arrivals the approximation satisfies the well-known pseudo-conservation laws and is exact in symmetric systems, vacation queues and general systems in LT. Moreover, the approximation gives exact results for systems with general renewal arrivals in the asymptotic regimes of HT or infinite switch-over times, as shown below. In the present section, comparable results are shown for higher moments of the distributional approximation.

Heavy-traffic. By construction, the distributional approximation is exact in systems with general renewal arrivals in HT. This property is very desirable from a practical perspective, since the proper operation of a system is particularly critical when the system is heavily loaded.

Large switch-over times. In case of deterministic switch-over times, the waiting time is only dependent on the total switch-over time in a cycle S, rather than the marginal switch-over times S_i (cf. [7]). A strong conjecture is presented in [19] that in this case the distribution of $\frac{W_i}{S}$ tends to a uniform distribution on $[0, \frac{1-\rho_i}{1-\rho}]$ as $S \to \infty$; for the case of Poisson arrivals, a rigorous proof of this result was given in [10]. It turns out that the distributional approximation as presented satisfies this result. To this end, consider the k^{th} moment of $\frac{W_{i,app}}{S}$, $k > 0$ as $S \to \infty$. It can easily be verified that $\lim_{S\to\infty} \mathbb{E}[\frac{W_{i,Boon}}{S}] = \frac{1-\rho_i}{2(1-\rho)}$, and hence,

$$\begin{aligned} \lim_{S\to\infty} \mathbb{E}\left[\left(\frac{W_{i,app}}{S}\right)^k\right] &= \frac{1}{k+1}\left(\frac{1-\rho_i}{1-\rho}\right)^k \lim_{S\to\infty} \prod_{i=1}^{k} \frac{2S\delta + i\sigma^2}{2S\delta + \sigma^2} \\ &= \frac{1}{k+1}\left(\frac{1-\rho_i}{1-\rho}\right)^k. \end{aligned} \tag{9}$$

This expression exactly coincides with the finite k^{th} moment of a uniformly distributed random variable Y on $[0, \frac{1-\rho_i}{1-\rho}]$. Thus, the k^{th} moment of $\frac{W_{i,app}}{S}$ converges to the k^{th} moment of Y when S tends to infinity, $k \geq 1$. Under certain conditions (which are met here), this moment-wise convergence implies convergence in distribution (cf. [4], Theorem 4.5.5). Therefore, the distributional approximation becomes exact in the case of deterministic switch-over times that tend to infinity.

5. Simulation study

In this section, we evaluate the accuracy of the approximation of the waiting-time distribution as presented in Section 3. First, we regard a rather arbitrary polling system and see how well the approximated and exact density functions coincide. The "exact" density function is determined by means of simulation. Then, we study the accuracy of the approximation in a wide range of parameter combinations by applying the approximation to a test

bed containing 10368 polling systems and summarising the errors of standard deviation and percentile approximations. Again, the "exact" standard deviations and percentiles are determined by means of simulation. Also in case of Poisson arrivals, where numerical methods exist to determine the exact distribution, we opt for simulation, since the determination of the exact values using numerical methods can be very cumbersome. All simulation results presented in this section are an average taken from a variable number of simulation runs with a length of at least 1,000,000 time units, such that the width of the confidence interval of the average is less than 1% of the value of the actual average.

5.1. Accuracy of the approximated density function

We consider a symmetric polling system with five queues. The load ρ equals 0.7, the SCV of the interarrival times at each queue are 0.25. All the service times and switch-over times are exponentially distributed with mean 1. Since there is no exact closed-form expression available for the waiting-time distribution in this case, we compare the density function of the approximated distribution with the simulated density function for an arbitrary queue. To obtain the latter, a kernel estimation was made based on a huge set of simulated waiting-time realisations. Both the approximated density function and the simulated density function are depicted in Figure 1. For the interarrival times, a gamma distribution was used with shape parameter 4 and inverse scale parameter 16.

Figure 1 shows that the shape of the exact waiting-time distribution closely resembles the approximation, which suggests that the approximation is useful for approximating the waiting-time distribution. The next subsection will show that the approximation works well not only in this case, but also in a variety of other polling systems.

5.2. Accuracy of approximated percentiles and standard deviation

In this subsection we assess the accuracy of the approximation by evaluating errors in the approximation of the standard deviation and several percentiles. We regard the standard deviation and several percentiles of the approximated distribution and the exact distribution of the waiting-time of the first queue in a large number of polling systems with exhaustive, cyclic service. The standard deviation and percentiles of the exact distribution are

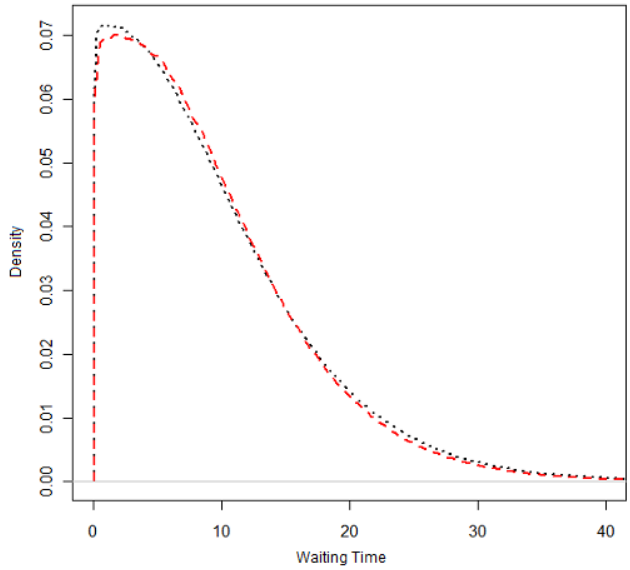

Figure 1: Approximated and simulated density function of the waiting-time of an arbitrary queue in the example in Subsection 5.1.

determined by means of simulation. We first give a general impression of the accuracy, after which we try to explore the impact of each of the parameters on the accuracy.

The parameter values contained in the test bed can be found in Table 1. There is no difference between the queues within a particular polling system in terms of service times and switch-over times. All parameters are explained above, except for the last one displayed in the table. $Q_2, \ldots, Q_N$ take on the same amount of load each. p_1 denotes what amount of load is taken on by Q_1 relative to the other queues. For example, if the first queue takes half, twice or five times as much load as any other, p_1 becomes 0.5, 2 or 5, respectively. In case of a symmetric system, $p_1 = 1$. In the test bed, several SCV's are regarded for several random variables. When the SCV equals one, an exponential distribution is fitted to the mean. For other SCV values, distributions commonly used in a two-moment fit were fitted to the first two moments; a mixture of two Erlang distributions in case of a SCV smaller than one and a H_2 distribution (with balanced means) for a SCV larger than one (cf. [16]).

For each polling system, the approximation error of the standard devi-

Notation	Parameter	Considered parameter values
N	Number of queues	$\{5, 10, 20\}$
ρ	Load	$\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$
$c_{A_i}^2$	SCV interarrival times	$\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$
$\mathbb{E}[B_i]$	Mean service times	$\{1\}$
$c_{B_i}^2$	SCV service times	$\{0, 1, 4\}$
$\mathbb{E}[S_i]$	Mean switch-over times	$\{0.2, 1, 10\}$
$c_{S_i}^2$	SCV switch-over times	$\{0, 1\}$
p_1	Measure of asymmetry	$\{0.5, 1, 2, 5\}$

Table 1: Parameter values of the test bed used in subsection 5.2.

ation and the approximation error of the 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th} and 95^{th}-percentiles are calculated. The errors are measured in a percentual absolute relative way, i.e.,

$$\Delta\% = \frac{|a - s|}{s} \times 100\%, \tag{10}$$

where a denotes the approximated value by means of the distributional approximation as presented in Section 3 and s denotes the exact value, determined by means of simulation.

Tables 2 and 3 show the errors made in approximating the standard deviation and in approximating the percentiles respectively in bins of 5%. This table suggests that the approximation performs best in case of $p_1 \leq 1$. Regardless, it is shown that the majority of the standard deviation approximations, and even more so the percentile approximation have errors lower than 5%.

To assess the impact of the system parameters on the performance of the approximation, the mean absolute relative errors of the standard deviation approximation and the percentile approximation are given in Tables 4 and 5, respectively, vertically categorised in the different rates of asymmetry, horizontally categorised in each of the relevant system parameters. Tables 4(a) and 5(a) show that the distributional approximation becomes better when N increases. The same behaviour of the approximation error in N is present in the approximations of [2, 12]. Tables 4(b) and 5(b) show a surprising effect of the SCV of the service times on the performance of the approximation.

p_1	Bins				
	0-5%	5-10%	10-15%	15-20%	20%+
0.5	69.98%	19.68%	5.02%	2.28%	3.05%
1	69.87%	19.33%	5.56%	2.39%	2.85%
2	67.67%	19.98%	6.17%	3.20%	2.97%
5	58.72%	22.07%	8.60%	4.86%	5.75%

Table 2: Mean standard deviation error of the approximation applied to the test bed, categorised in bins of 5%.

p_1	Bins				
	0-5%	5-10%	10-15%	15-20%	20%+
0.5	82.25%	10.77%	3.26%	1.07%	2.65%
1	82.14%	10.98%	3.44%	1.03%	2.41%
2	82.25%	11.17%	3.38%	1.04%	2.16%
5	77.63%	13.32%	4.76%	2.00%	2.29%

Table 3: Mean percentile error of the approximation applied to the test bed, categorised in bins of 5%.

While in case of $c^2_{B_i} = 0$ the standard deviation approximation error seems to grow with p_1, the same effect does not seem to happen when $c^2_{B_i} = 4$. Also, if $c^2_{B_i} = 1$ or $c^2_{B_i} = 4$, the approximation error of the standard deviation and the approximation error of the percentiles do not seem to react in the same way to changes in p_1. Tables 4(c) and 5(c) suggest that the distributional approximation becomes better when the variance of the switch-over times increases. Tables 4(d) and 5(d) suggest that approximations become better as ρ approaches 1, i.e. as the system becomes closer to HT. This is very plausible, since by construction the approximation is exact in HT, as discussed in Section 4. According to Tables 4(e) and 5(e) the approximations seem to become better as the switch-over times larger. This is in line with the observation made in Section 4 that the approximations becomes exact as the total switch-over time tends to infinity. Also, using (6) one can show that the moments of the distributional approximation become less dependent of σ^2 and δ as switch-over times become smaller, which gives support to the plausibility of the approximation becoming less reliable when the switch-over times become relatively small. Finally, both Tables 4(f) and 5(f) show that the approximations' quality is dependent on the SCV of the interarrival times, but again an interaction effect with the value of p_1 is observed.

(a)

p_1	N		
	5	10	20
0.5	5.97	4.02	3.17
1	5.99	4.06	3.25
2	6.41	4.31	3.29
5	9.08	5.76	3.77

(b)

p_1	$c^2_{B_i}$		
	0	1	4
0.5	3.91	3.25	5.99
1	4.15	3.33	5.81
2	4.80	3.64	5.57
5	6.97	5.49	6.15

(c)

p_1	$c^2_{S_i}$	
	0	1
0.5	4.49	4.28
1	4.59	4.27
2	8.80	4.46
5	6.73	5.68

(d)

p_1	ρ					
	0.5	0.6	0.7	0.8	0.9	0.95
0.5	8.22	6.66	5.11	3.48	1.81	1.06
1	8.37	6.77	5.12	3.49	1.79	1.06
2	8.82	7.14	5.42	3.65	1.87	1.12
5	11.02	9.30	7.36	5.19	2.75	1.60

(e)

p_1	$\mathbb{E}[S_i]$		
	0.2	1	10
0.5	7.35	3.37	2.44
1	7.34	3.53	2.33
2	7.44	4.07	2.30
5	8.76	6.79	3.06

(f)

p_1	$c^2_{A_i}$							
	0.25	0.5	0.75	1	1.25	1.50	1.75	2
0.5	4.11	4.13	3.95	4.04	4.34	4.59	4.82	5.10
1	4.10	4.14	4.00	4.06	4.40	4.67	4.91	5.18
2	4.18	4.42	4.28	4.33	4.67	4.93	5.18	5.37
5	4.77	5.20	5.28	6.00	6.48	6.94	7.30	7.65

Table 4: Mean standard deviation error categorised by the value of p_1 vertically and the number of queues (a), the SCV of the service times (b), the SCV of the switch-over times (c), the total load (d), the mean switch-over time (e), and the SCV of the interarrival times (f) horizontally.

Table 6 shows the mean absolute relative error categorised per tested percentile. Generally, the 80% percentiles seem to be approximated best.

From the test-bed results we can conclude that the approximation performs well over a wide range of parameter combinations. In case of extremely variable service times, low load and negligibly small switch-over times, the relative error becomes worse. The worst-case scenarios found in the testbed

(a)

p_1	N		
	5	10	20
0.5	5.77	3.27	1.99
1	5.45	3.23	2.05
2	5.10	3.18	2.02
5	5.87	3.50	2.15

(b)

p_1	$c^2_{B_i}$		
	0	1	4
0.5	2.15	2.40	6.48
1	2.28	2.27	6.19
2	2.44	2.31	5.56
5	3.35	3.03	5.14

(c)

p_1	$c^2_{S_i}$	
	0	1
0.5	3.87	3.49
1	3.77	3.39
2	3.58	3.29
5	4.14	3.54

(d)

p_1	ρ					
	0.5	0.6	0.7	0.8	0.9	0.95
0.5	5.99	5.79	4.46	2.89	1.37	0.79
1	5.86	5.65	4.28	2.75	1.30	0.79
2	5.67	5.44	4.05	2.56	1.22	0.75
5	6.14	5.86	4.60	3.10	1.55	0.91

(e)

p_1	$\mathbb{E}[S_i]$		
	0.2	1	10
0.5	7.50	2.19	1.34
1	6.98	2.31	1.45
2	6.43	2.42	1.46
5	6.14	3.64	1.74

(f)

p_1	$c^2_{A_i}$							
	0.25	0.5	0.75	1	1.25	1.5	1.75	2
0.5	4.20	3.59	3.25	3.31	3.48	3.65	3.86	4.09
1	3.97	3.41	3.14	3.29	3.44	3.61	3.78	3.97
2	3.53	3.13	3.04	3.30	3.45	3.56	3.67	3.80
5	3.15	3.26	3.52	3.80	4.00	4.18	4.32	4.50

Table 5: Mean percentile error categorised by the value of p_1 vertically and the number of queues (a), the SCV of the service times (b), the SCV of the switch-over times (c), the total load (d), the mean switch-over time (e), and the SCV of the interarrival times (f) horizontally.

in terms of absolute relative error are approximations of the 50^{th} percentile in systems with $N = 5$, $\rho = 0.5$, $c^2_{B_i} = 4$ and $\mathbb{E}[S_i] = 0.2$ having errors with an order of magnitude of 100%. However, in practice these characteristics are uncommon. For example, in production systems settings like $c^2_{B_i} = 4$ are hardly found due to the just-in-time philosophy, which dictates to reduce variability in e.g. service times in order to reduce in-process inventory. Also, these systems are typically utilized beyond $\rho = 0.5$ to increase productivity, and switch-over periods are commonly longer than service periods. More-

p_1	Percentile						
	40^{th}	50^{th}	60^{th}	70^{th}	80^{th}	90^{th}	95^{th}
0.5	5.90	5.23	4.15	2.88	1.72	2.29	3.57
1	5.73	5.00	3.93	2.71	1.67	2.37	3.63
2	5.42	4.59	3.60	2.50	1.67	2.49	3.78
5	6.29	4.87	3.40	2.11	1.88	3.42	4.92

Table 6: Mean absolute relative errors categorised in the several percentiles.

over, in case of a low load and small switch-over times, although the relative error of the percentile approximations can be high, the absolute errors may still be rather small when compared to the order of longitude of service time durations. Therefore, the sojourn time distribution is already much better approximated in these situations.

6. Further research

The present paper gives birth to a variety of directions for further research. Firstly, the distributional approximation for cyclic systems with exhaustive service may be generalised to models with branching-type service policies [14], non-cyclic periodic server routing [13] and other model variations. Secondly, the simple closed-form expression may act as a basis for design decisions within polling systems. Finally, one could attempt to improve the approximation by deriving an interpolation approximation for higher moments of the waiting-time and, subsequently, fit a phase-type distribution. Such an extension would be of particular interest if one has more information than the first two moments of the interarrival, service and switch-over time distributions (see also Remark 2). However, this would impel one to considerably extend the analysis of [2], while potentially losing the simple form of the current distributional approximation.

Acknowledgements

The authors would like to thank Marko Boon for placing parts of his polling system simulation program at their disposal, and for helpful comments on earlier drafts of this paper.

References

[1] J.P.C. Blanc (1992). Performance evaluation of polling systems by means of the power-series algorithm. Annals of Operations Research 35, 155-186.

[2] M.A.A. Boon, E.M.M. Winands, I.J.B.F. Adan and A.C.C. van Wijk (2009). Closed-form waiting-time approximations for polling systems. Eurandom Report No. 2009-030, Eindhoven.

[3] G. Choudhury and W. Whitt (1994). Computing transient and steady state distributions in polling models by numerical transform inversion. Performance Evaluation 25, 267-292.

[4] K.L. Chung (1974). A Course in Probability, 2nd ed. Academic Press, New York.

[5] D. Down (1998). On the stability of polling models with multiple servers. Journal of Applied Probability 35, 925-935.

[6] D. Everitt (1986). Simple approximations for token rings. IEEE Transactions on Communications 34, 719-721.

[7] S.W. Fuhrmann (1992). A decomposition result for a class of polling models. Queueing systems 11, 109-120.

[8] K.K. Leung (1991). Cyclic-service systems with probabilistically-limited service. IEEE Journal on Selected Areas in Communications 9, 185-193.

[9] H. Levy and M. Sidi (1990). Polling systems: applications, modeling, and optimization. IEEE Transactions on Communications 38, 1750-1760.

[10] R.D. van der Mei (1999). Delay in polling systems with large switch-over times. Journal of Applied Probability 36, 232-243.

[11] R.D. van der Mei and E.M.M. Winands (2008). A note on polling models with renewal arrivals and nonzero switch-over times. Operation Research Letters 36, 500-505.

[12] T.L. Olsen (2001). Approximations for the waiting-time distribution in polling models with and without state-dependent setups. Operations Research Letters 28, 113-123.

[13] T.L. Olsen and R.D. van der Mei (2005). Polling systems with periodic server routing in heavy-traffic: renewal arrivals. Operations Research Letters 33, 17-25.

[14] J.A.C. Resing (1993). Polling systems and multitype branching processes. Queueing Systems 13, 409-426.

[15] H. Takagi (1985). Analysis of polling systems. MIT Press, Cambridge.

[16] H.C. Tijms (1994). Stochastic models: An algorithmic approach. Wiley, Chichester.

[17] V.M. Vishnevskii and O.V. Semenova (2006). Mathematical methods to study the polling systems. Automation and Remote Control 67(2), 173-220.

[18] W. Whitt (1982). Refining diffusion approximations for queues. Operations Research Letters 1, 165-169.

[19] E.M.M. Winands (2009). Branching-type polling systems with large setups. To appear in OR Spectrum.